DOCUMENT RESUME

ED 221 563 TM 820 585

TITLE	Development of the State Speaking Assessment Instrument: Reliability and Bias Study, Summary Report. Basic Skills Improvement Policy.
INSTITUTION	Massachusetts State Dept. of Education, Boston. Bureau of Research and Assessment.
PUB DATE	Jun 82
NOTE	48p.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Achievement Tests; Basic Skills; Elementary Secondary Education; Policy Formation; *Speech Evaluation; State Boards of Education; State Programs; *Test Bias; Test Construction; Testing Programs; *Test Reliability
IDENTIFIERS	*Basic Skills Improvement Policy (Massachusetts)

ABSTRACT

Since the approval of the Basic Skills Improvement Policy in 1978, the Massachusetts Department of Education has been developing tests and alternative forms for the assessment of student achievement in five basic skills content areas: reading, writing, mathematics, listening, and speaking. Because of the lack of previous research on which to draw in the assessment of speaking skills, the Department previously determined the method by which speaking assessment would be done, and had assessed a large sample of twelfth grade students around Massachusetts using this assessment procedure. This year's study was to investigate reliability and bias issues related to the instrument and to produce final materials for the training of raters to assess speaking. The results demonstrated that there is virtually no bias in the instrument or in the ratings. Raters of all ethnic groups and both sexes exhibited virtually the same pattern of ratings. The recommendation of the study is to finish this work by determining methods to minimize the cost and inconvenience of assessing speaking skills, and to produce a sufficient quantity of materials for use by all districts. Study results relating to rating differences and rater agreement, and statistics describing rater workshop performance in the field are appended. (Author/PN)

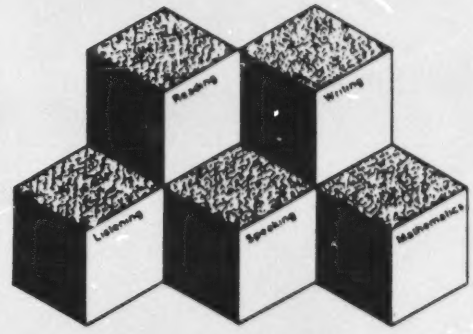
****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221563

Basic Skills Improvement Policy

Development of the State Speaking Assessment Instrument:

Reliability and Bias Study



U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. DiBella

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SUMMARY REPORT

TM 820 585

Massachusetts Department of Education

June, 1982

Massachusetts Board of Education

Donald R. Walker, Chairperson
Howard A. Greis, Vice Chairperson

John W. Bond
Millie R. Clements
Joseph P. Crimmins
James R. Grande
James L. Green
Charles T. Grigsby
Armando Martinez
Edwin M. Rossman
Joseph C. Savery
Dorothea A. Zanetti

John H. Lawson, Commissioner of Education

John B. Duff, Ex Officio, Chancellor, Board of Regents

Bureau of Research and Assessment

Allan S. Hartman, Director
Matthew H. Towle
Richards S. Williams
Dorothy Earle
Betty Hancock

The Massachusetts Department of Education insures equal employment/educational opportunities/and affirmative action, regardless of race, color, creed, national origin, or sex, in compliance with Title IX, or handicap, in compliance with Section 504.

Publication #12823 approved by John Manton, State Purchasing Agent

ACKNOWLEDGMENTS

The preparation of this report benefited from the valuable advice and assistance of many people. Special acknowledgments are due the following:

Study Review Committee

Janet Doe
Joanne Gurry
Anne Keenan
Eileen Malloy
Pamela Mason
Jane Sherman

Advisory Committee on Basic Skills Improvement

Fred Andelman
Christine Angello
Eugene Applebaum
Claudette Chagnon
Thomas Deveney
Joseph DiFabio
Janet Doe
Roberta Doering
Maxine Edmunds
Abel Fidalgo
Tassos Filledes
Melissa Fleming
Edna Frank
Donald Frizzle
Paul Gauvreau
Richard Granger
Joanne Gurry
Donald Hayes
Eugene Kirby
Mary Lewis
Jeannette McInnes
Eileen Malloy
Pamela Mason
Gertrude Moskowitz
Neil Mosman
Donald Palko
John Reynolds
Walter Secatore
Vincent Silluzio
Florence Sotiros
L. Stephen Traw

BASIC SKILLS IMPROVEMENT POLICY

DEVELOPMENT OF THE

STATE SPEAKING ASSESSMENT INSTRUMENT:

RELIABILITY AND BIAS STUDY

JUNE, 1982

Bureau of Research and Assessment
Massachusetts Department of Education
Quincy Center Plaza
1385 Hancock Street
Quincy, Massachusetts 02169

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY 1

II. BACKGROUND 2
1979-80 4
1980-81 5
Plans for 1981-82 6
Speaking Assessment Instrument 7

III IMPLEMENTATION OF THE STUDY 9
Revision of the Workshop 9
Selection of the Workshop Participants 10
Conduct of the Workshop and Debriefing 11
Collection of Data 12
Analysis of Data 12

IV. RESULTS 13
A. Field Data 13
B. Efforts to Use the Workshop Examination Data ... 15
C. Bias 16
D. Revising the Workshop and Instrument 17
E. Estimates of Cost 20

V. DISCUSSION 23
Critical Technical Issues 24
Consistency of Mastery Decisions 25
Use of Multiple Raters 26
Remaining Sources of Unreliability 27
Efficiencies of Individual Administration 28

VI. CONCLUSIONS AND RECOMMENDATIONS 29

APPENDICES

A. Study Results Related to Sex and Ethnic Group Differences in Ratings i

B. Study Results Related to Agreement of Raters in the Field iii

C. Statistics Describing ther Performance of Raters in the Workshop and in the Field vi

I. EXECUTIVE SUMMARY

Since the approval of the Basic Skills Improvement Policy in 1978, the Department has been developing tests and alternative forms for the assessment of student achievement in five Basic Skills content areas: reading, writing, mathematics, listening, and speaking. Because of the lack of previous research on which to draw in the assessment of speaking skills, the Department has been careful and methodical in its development of a test for the assessment of speaking skills. This year's study provided some major elements that were needed before moving into the final stages leading to full-scale implementation.

In two previous years of study, the Department had determined the method by which speaking assessment would be done, and had assessed a large sample of twelfth grade students around Massachusetts using this assessment procedure. The purpose of this year's study was to investigate reliability and bias issues related to the instrument and to produce final materials for the training of raters to assess speaking.

In order to implement the study, 51 participants from 21 schools throughout Massachusetts were trained to assess speaking skills at a two-day workshop. Following the workshop, participants took an exam designed to evaluate their skills in conducting the assessment and then returned to their schools to assess 50 students. Participants worked as teams so that the agreement between pairs of raters could be examined.

The results of the study showed that the workshop was highly successful, but several areas for further improvement were identified both by project staff and the participants. These improvements are being incorporated into revisions of materials that will be provided to the Department before completion of the study. While many of these revisions are minor, it is believed that some of them will yield a marked improvement in rater accuracy, and that use of these materials in the future will lead to even higher levels of agreement between raters than those found this year.

The results also demonstrated that there is virtually no bias in the instrument or in the ratings. Students of all ethnic groups and both sexes exhibited the same pattern of scores on all tasks and all scoring dimensions in the test. Also, the workshop exam, which included students of all three major ethnic groups (black, Hispanic, and white) and both sexes, raters of all ethnic groups and both sexes exhibited virtually the same pattern of ratings.

The levels of agreement between raters on the field ratings were not as high as expected, but were high enough so that for at least half the students tested, one rating would be sufficient to accurately assess the student as being above or below the district standard. For the remaining students, some additional information would need to be gathered before making a final determination about them. Almost

certainly, a valuable source of additional information for them would be retesting using a second rater.

Contrary to expectations, the workshop examination did not consistently predict which raters would agree most closely in the field. Consequently, improvement in the degree of agreement between raters cannot come from eliminating certain raters, but must come from the improvements in the training materials noted above.

At this point, development of the speaking instrument is virtually complete. The recommendation of the study is to finish this work by determining methods to minimize the cost and inconvenience of assessing speaking skills, and to produce a sufficient quantity of materials for use by all districts. The study would complete all developmental efforts in time for review by the State Board in June, 1983.

II. BACKGROUND

In August, 1978, the Massachusetts State Board of Education approved a document entitled Policy on Basic Skills Improvement. The Policy, which was followed by Regulations approved by the Board of Education in January, 1979, charts a course for public schools in the Commonwealth to improve the teaching and learning of basic skills.

The Policy requires all schools to develop plans to evaluate basic skills competency. These plans must incorporate the basic skills objectives for reading, writing, mathematics, listening, and speaking that are specified in the Regulations.

Schools must evaluate each student's achievement of these objectives, and, at secondary level, must choose one or more of the following evaluation instruments to evaluate such achievement:

1. evaluation instruments produced by and available from the Department of Education,

2. commercially available instruments approved by the Department of Education, or

3. locally developed evaluation instruments approved by the Department of Education as being comparable to one of the above.

In practice, of course, one administration of any one instrument does not provide infallible measurement. Consequently, the best implementations of the Policy incorporate more than the results of one administration of a test in their evaluation of a student. Typically, students first take the test; those who have scores near the standard established by the district then are reevaluated--sometimes by a readministration of the test, and sometimes by other evidences of competency that the district can bring to bear.

Since the approval of the Regulations, the Department of Education has been developing and evaluating tests in the five specified content areas. While the timetables for the developmental efforts in the five areas have differed, each has followed the same basic pattern:

- development of an initial instrument
- administration of that instrument to a representative sample of twelfth grade students
- an analysis of reliability and bias issues
- development of alternative forms
- an investigation of feasibility issues
- wide-scale implementation of the instrument and alternative forms.

As opposed to some of the other content areas, very little was available to draw upon when the Department began its development of a speaking test. Therefore, the timetable for implementation has been lengthened somewhat for speaking. However, the development of a speaking instrument

has followed the same basic pattern as the other content areas.

The narrative below summarizes the progress that was made during the first two years of efforts to determine a practical, technically sound process for measuring speaking. For further detail, the reader is referred to those two years of study. Copies of these reports are available from the Department of Education.

1979-80

As noted above, determination of whether students have surpassed the standard of minimum competence established by their district is ideally a multi-stage decision process. That is, one administration of an instrument (ideally inexpensive and easy to administer) separates out those students who are far from the standard. Then, those who are "too close to call" are reassessed. In keeping with this thinking, the Department's initial efforts to develop a speaking instrument included a two-staged approach for assessing speaking skills. This approach included:

- Teacher observation ratings
- One-on-one ratings

The teacher observation ratings were to be used as a general measure of a student's speaking performance. In this approach, two teachers who had a student currently enrolled in their classes were to independently rate the student's general speaking performance using a speaking rating scale. Teachers would base their ratings of a student on observations of the student's performance in normal classroom activities, such as asking questions, responding to questions, explaining how to do something, giving a report to the class or talking with other students in discussion groups. The teachers were to consider the student's performance on the average since the beginning of the current semester. It was hoped that the teacher observations could be used as an inexpensive screening device. Then, for those students whose observation scores were too close to the district standard, a one-on-one rating would be conducted.

The one-on-one ratings were a "focused measure" of a student's speaking performance. (A focused measure is an occasion for observing speech where that is the principal reason for observing the student, as opposed to, for example, a classroom setting where the principal reason for observing students is to observe their progress in the class, not their speech.) In this approach, one specially trained rater rated the student's performance using similar speaking rating scales to those used in the teacher observation approach. However, instead of basing ratings on classroom observations, the rater gave the student four specific speaking tasks and rated the student's responses to these tasks.

The two-staged approach was field tested during the Spring of 1980 with a random sample of 12th graders from 49 schools across the state. The results of the study were promising and provided a statewide overview of students' speaking skills. However, the conclusion of that study was that several elements were needed before attempting full-scale implementation:

- a need for some changes in the assessment instruments, particularly a need for increased variability in both the teacher observation and the one-on-one ratings;

- a need for detailed training materials so that the assessments could be implemented easily in schools in the future; and,

- a need to study the issue of reliability of these instruments in more detail.

1980-1981

In the development of any instrument, a major concern is reliability --will the score obtained for an individual (or the decision made about that individual) be the same if the individual is measured a second time? The only real test of reliability comes in replication. If scores are obtained on individuals, and then, upon replication, a second set of scores is found to closely match the first set, the instrument has high reliability. Therefore, the 1980-81 study was conducted to determine the reliability and feasibility of the two-staged approach in typical schools. The tasks of the study were to refine the teacher observation and one-on-one instruments, to develop appropriate material to train the raters and to conduct trial assessments of eighth graders in several representative schools. Eighth graders were selected for study because, at that time, many districts were conducting basic skills assessment at that level.

In order to test the reliability of the assessment instrument, a "replication study" was designed. In this replication study, all of the assessment ratings were duplicated. Instead of conducting two teacher observations for each student, four teacher observations were obtained; instead of conducting a single one-on-one rating, two one-on-one ratings were obtained. The replicated information was used to measure the reliability of the instruments. The observation rating of two of the teachers were compared to ratings of the other two teachers to see if both pairs of teachers rated the student in the same way. Similarly, the one-on-one ratings of one rater were compared to the one-on-one ratings of the other rater to see if both raters rated the student in the same way.

The field testing was conducted in eight schools representing a variety of school settings. Three schools were located in large cities with minority student populations, three in suburban areas and the remaining two in primarily rural areas. Within each school, 60 eighth grade students were selected at random to participate. Each student received observation ratings from four of his or her teachers. In addition, each student received one-on-one ratings from two trained raters within the school district who had received training at a workshop conducted by the project staff.

The results of this study provided information that led the Department to revise its approach toward the assessment of speaking skills.

One primary finding was serious difficulties with the teacher observation ratings. Teachers disagreed often in their ratings of the speaking skills of students. As a consequence, too few students were classified as having definitely passed or failed as a result of the teacher judgments. It was surmised that the teachers were observing students under such varied conditions of speaking that the teachers were coming to substantially different conclusions about each student.

It is logical to assume that students have more opportunity to display their speaking skills in some types of classes than others. However, the level of agreement among teachers of all subject matters was about the same.

While the results of the teacher observations were discouraging, the results of the one-on-one ratings were good. There was substantially more agreement about the speaking skills of students when they were rated under standardized conditions. An additional finding that influenced the next stage of development was that while the ratings evidenced substantial agreement, pairs that agreed closely in their ratings on a short examination given at the end of the training workshop agreed most closely on students when rating in the field. This result led to an interest in determining whether a longer and more carefully constructed examination would predict which raters would agree best in the field.

Plans for 1981-82

As a consequence of what was learned during the first two years of developing a speaking assessment instrument, in August, 1981, the Board of Education approved a motion to continue efforts to develop such an instrument. The motion effectively eliminated teacher observations as a possible method for assessing speaking and directed the Department to focus its efforts on continued development of the one-on-one ratings.

In response to this motion, the Bureau of Research and Assessment contracted with RMC Research Corporation to conduct a study that would include the following:

1. revision of the workshop used to train raters.

2. training of raters from a wide variety of schools.

3. determination of whether the sex or ethnic group of either students or raters is a factor in assigning ratings.

4. determination of whether two raters hearing the same student speaking would assign similar scores to that student (provide reliable ratings).

The Speaking Assessment Instrument

The instrument used to assess speaking skills in the one-on-one rating situation consists of four tasks which are presented to students orally one at a time. The four tasks represent four categories of speaking listed in the Basic Skills Objectives. Those categories, and a summary of the tasks used within the category this year, are as follows:

1. Description -- describe a favorite class or extracurricular activity.
2. Emergency -- call the fire department to report smoke in the house.
3. Sequence -- describe how to cook something.
4. Persuasion -- convince the principal to make one change in the school.

After a student responds to a task, a trained rater scores the student on each of the following scoring dimensions:

1. Delivery, which is concerned with how a message is transmitted (i.e., a speaker's volume, rate and articulation);
2. Language, which is concerned with the grammar and words used by the speaker;
3. Content, which is concerned with the amount and relevance of information transmitted; and,
4. Organization, which is concerned with how the speaker orders or links ideas communicated.

Each of these dimensions is scored on a scale from 1-5:

1. Many problems with speaking performance
2. Some problems with speaking performance
3. Absence of problems with speaking performance and absence of characteristics of good speaking performance
4. Some characteristics of good speaking performance
5. Many characteristics of good speaking performance

Thus, the total score for any one task can range from 4 to 20 points; for the entire test (four tasks), total scores can range from 16 to 80 points.

The two critical elements of the rating process, and thus the major concerns of the study conducted this year, are first, the tasks themselves:

1. Do they perform equally well for students regardless of sex or race?
2. Do they stimulate students to speak and subsequently demonstrate to their best advantage the students' speaking skills?

Second, the ratings assigned by raters:

1. Do two raters assign the same score to any given response?
2. Are the ratings independent of any bias because of the sex or race of the students or raters?

III. IMPLEMENTATION OF THE STUDY

The study consisted of the revision of the workshop, selection of workshop participants, conduct of the workshop, collection of data and analysis of data. The following sections will describe the implementation of each of these activities.

A. Revision of the Workshop

A workshop to train raters to conduct one-on-one ratings had been used in the 1980-81 study. As would be true with any first effort, it was felt that revisions would improve it. Initial efforts in the project, therefore, were oriented toward making the following changes in the workshop:

1. The structure of the workshop was drastically altered. Instruction on how to score each task and dimension was systematically presented. The new format was as follows:

   a. For the first task presented (description), workshop participants were instructed in scoring the delivery dimension. After instruction in this area was complete, participants practiced assigning scores in delivery after listening to the taped responses to the description tasks given by four students.

   b. After instruction and practice on the delivery dimension were completed, the same routine was used to present, in order, the dimensions of language, content, and organization.

   c. As final practice in scoring the description task, eight final samples were played for participants. The first four were played twice, as the participants rated two scoring dimensions at a time. The last four were played only one time. It was this final exercise -- rating all four dimensions simultaneously -- that most closely matched the actual skills needed by raters in the field.

   d. After all this training and practice was completed on the description task, the same instructional pattern was used to train the participants to score, in order, the emergency, sequence, and persuasion tasks.

2. Because of the vast increase in new training materials presented to the participants, the time for the workshop was expanded from one day to two.

3. A far larger number of examples were played for participants.

4. A complete high fidelity sound system with multiple speakers was introduced to play the tapes. Use of such equipment greatly facilitated hearing finer distinctions among examples.

5. Transcripts of many tapes and a revision of the Test Administrator's Manual were provided in the training manual.

6. Participants were encouraged to comment on the proposed rating structure, and minor revisions were made during the workshop based on their feedback.

7. Participants were frequently divided into small groups to facilitate personal involvement in the process.

8. During these small group sessions, individual team leaders representing each of the three major ethnic groups facilitated discussion.

9. The overhead transparencies used were closely coordinated with the written materials presented to the participants.

10. The testing of participants conducted at the end of the workshop was expanded, and included presentation of students both on audiotape and videotape.

11. The workshop was evaluated by participants at its conclusion.

12. A debriefing, designed to obtain feedback from the participants on both the workshop and collection of data, was held 6 weeks after the workshop.

B. Selection of the Workshop Participants

A major purpose of the study was to determine whether the sex/ethnic group of a rater would have any influence on ratings when the sex/ethnic group of the student was changed. Therefore, the schools and people selected for the study were not chosen at random, but with an emphasis on balanced representation. The original goal was to select 42 participants equally divided on sex and ethnic group from 21 schools chosen from all areas of Massachusetts.

The initial step taken was to select the 21 schools. In order to assure that each school selected would contribute information on the performance of minority students, only those schools enrolling at least eight minority eighth grade students were considered. Approval for participation of the 21 selected schools was obtained from the respective superintendents and principals.

Once the schools were selected, two participants from each school were requested. Participants were not always teachers in the selected school; some were teachers at other schools in the districts or members of the district's cental office staff.

Although only two participants per district were requested, several districts elected to send three or four participants. The final total realized was 51 participants from 21 schools -- 13 blacks, 10 Hispanics, and 28 whites; 28 were female and 23 male.

C. Conduct of the Workshop and Debriefing

The workshop was conducted on January 20 and 21, 1982, at the Greater Boston Regional Education Center. All 51 participants attended in addition to three members of the Bureau of Research and Assessment, three instructors representing the contractor, and three members of advisory groups.

Participants were provided with manuals that provided an agenda, instructional material related to the workshop, rating sheets, an administrator's manual, and several miscellaneous aids to learning. After a short introduction and overview by the Department, instruction began.

The workshop proceeded very much according to plan, except that more time was needed for instruction and practice than had been planned. Participants became very involved in the process and very enthusiastically accepted instruction and provided feedback. The expanded time spent on instruction and practice led to a reduced amount of time spent on training in administration of the test than had been originally scheduled. Participants assisted by suggesting an alternative order of presentation on the second day that saved some time.

The workshop concluded the second day with the administration of an examination and an evaluation of the workshop. The purpose of the exam was to determine the overall averages for the total group of participants, to determine which raters matched most closely to these group averages, to find out whether bias existed in the ratings, and to identify those pairs which matched each other most closely. The exam was in two parts: an audiotape and a videotape. Both parts contained the responses of 12 students to one task; each of the four tasks was responded to by 3 students. The students on both parts consisted of 6 males and 6 females, and 4 blacks, 4 Hispanics and 4 whites.

As will be noted later in this report, the evaluations of the workshop provided by the participants were extremely positive. But for those familiar with the production of instruction of this type, perhaps the most accurate reflection of the success of the workshop is contained in the observation that all 51 participants attended every session of both days, and when the examination was given at the end of the second day, all 51 participants remained to take it.

The debriefing was held on March 18, again at the Greater Boston Regional Education Center. The purpose of this meeting was to gather feedback from the participants. The debriefing began with a readministration of the examination that had been given at the end of the workshop. Partly to satisfy the curiosity of the participants, the audiotape was played this second time as a videotape so that participants could see, as well as hear, the students they had rated at the workshop. After this second administration of the test, participants were told how they had scored each student the first time, as well as the group averages the first time. Next these results were compared and discussed in a roundtable format.

The debriefing consisted largely of a structured interview. Participants were divided into four groups, with a project or Department staff member serving as recorder and facilitator for each group. Participants were asked to comment on the workshop exam, the structure of the workshop, the workshop materials, the speaking tasks, and their experiences in administering the tests in the field.

D. Collection of Data

Three types of data were collected:

- opinions and cost estimates about the workshop and field collection process,
- workshop examination results,
- ratings in the field of students by teachers.

The first type of data included a questionnaire about the workshop completed by participants at its conclusion, notes provided by those who attended the debriefing, and notes of facilitators gathered during a structured interview conducted at the debriefing. The second type of data -- the workshop examination results -- were the ratings by the participants of the students in the audiotape and videotape described in the preceding section.

The main data collection effort was the field ratings conducted by the participants between the workshop and the debriefing. The team from each school (usually two raters, but three or four raters in some cases) examined and rated between 40 and 50 students in each school. The students were selected by RMC staff from lists supplied by the schools. As mentioned earlier, the students were selected to include equal numbers, insofar as possible, from each ethnic group (blacks, Hispanics, and whites). Fifty students were selected from each school as well as 10 alternates. Raters were instructed to test students on the first list where possible, but to substitute alternates when necessary.

Data were received back from all schools (although one provided data too late to be keypunched and included in the analyses). All raters but one had rated a full complement of students. The exception was not granted release time to conduct a second day of testing. Fortunately, this occurred in the one school where four raters were scoring each student; consequently, the partial data of this rater was deleted. Data were received for 1,014 students, all of whom were rated by two or three raters. Of the total sample of students, 252 (25%) were black and 245 (24%) were Hispanics.

E. Analysis of Data

The workshop and debriefing test data and the field data were all keypunched so that computerized analyses of the data could be conducted. Other data, such as the workshop evaluation data and the feedback from workshop participants, were collated manually. Details of the analyses and their results are provided in the next section of this report.

IV. RESULTS

The analyses of data and results are divided into five parts in this section:

- the field data,
- efforts to use the workshop exams to explain the field data,
- information on the bias of raters and the field instruments,
- information on revising the workshop,
- estimates of cost obtained from the raters.

A. The Field Data

At the conclusion of the workshop, the raters paired off and tested 50 students per pair in their schools. The data provided by this field testing provided evidence on the consistency of ratings assigned by different raters hearing the same student at the same time.

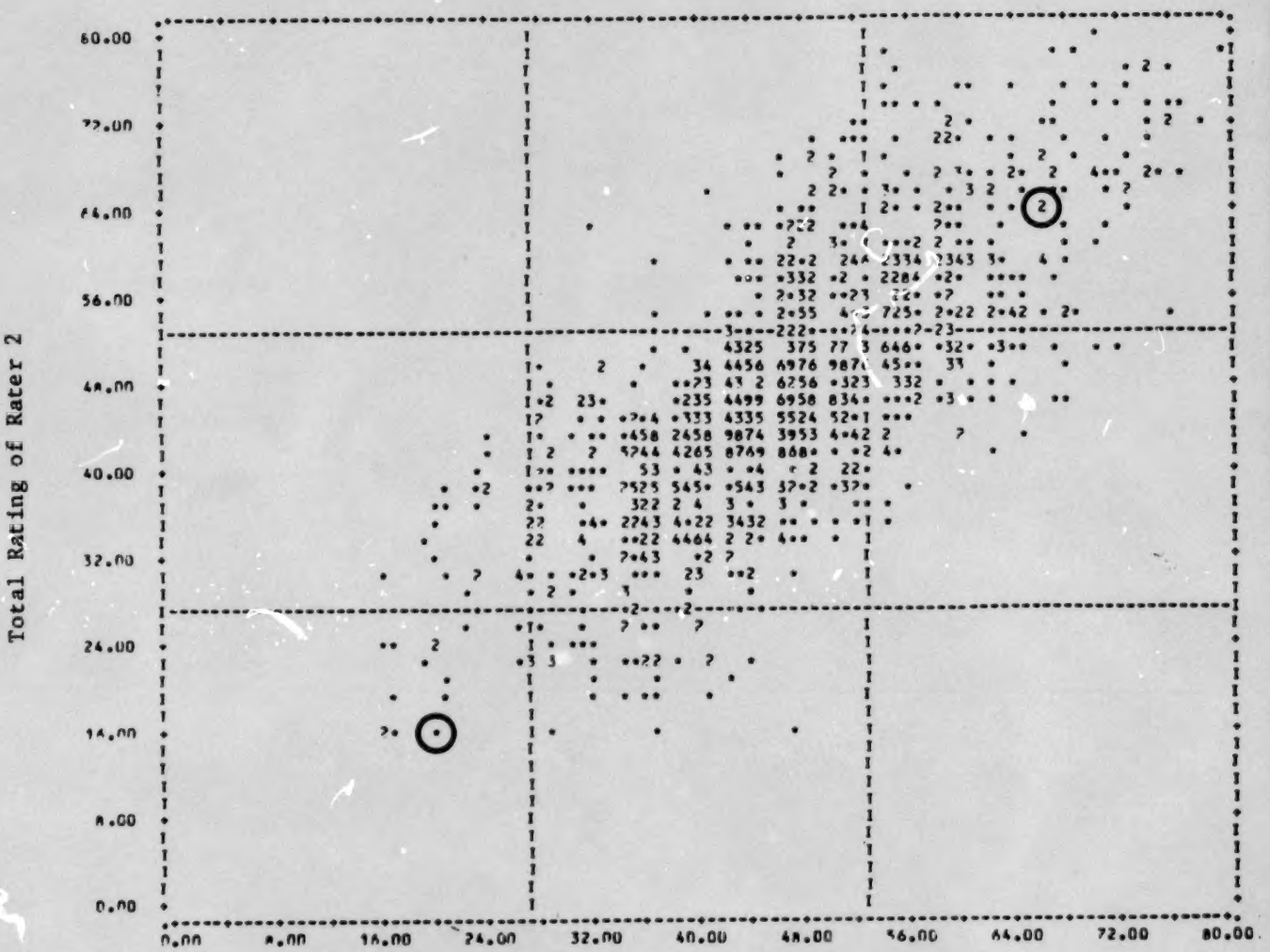
Raters were instructed to work as a team; one of them was to administer the speaking instrument, with the other rater(s) in that unit present. Raters then rated the student independently. Great emphasis in the workshop was placed on impressing the raters with the fact that they were not to reveal any scores to each other until all the ratings were finished. From feedback received at the debriefing, it is our belief that this, in fact, occurred and that the data received may be assumed to be independent estimates of student performance. This point is important to note, for if raters discussed their ratings with each other, the degree of agreement observed in this study would be an overestimate of the true agreement.

All schools conducted their ratings as agreed. The total amount of data received for inclusion in this report came from the testing of 1,014 eighth grade students: 501 males and 513 females; and 252 blacks, 245 Hispanics, and 517 whites. These students were all rated by two or three raters.

The first analysis done was to determine the amount of agreement between the raters. The results are displayed graphically as Figure 1. This figure depicts every pair of ratings assigned to students by the field raters. Two points have been circled to assist the reader in understanding and interpreting the figure. The asterisk circled in the lower left-hand corner of the graph depicts a student who was assigned a score of 20 by Rater 1 and a score of 16 by Rater 2. Keeping in mind that the range of possible scores for the test was 16 to 80, this was a student who was rated as very poor by both raters. The "2" circled in

Figure 1

Scatterplot of Total Scores Assigned to Students by Two Raters



the upper right-hand corner of the graph indicates that there were 2 students who were assigned a score of 66 by Rater 1 (not all score points are indicated on the graph, and some estimation of the intermediate score points is necessary in this display) and a score of 64 by Rater 2. Again, there was substantial agreement between the raters as to the speaking skills of these two students: both were rated very highly by both raters.

In standard statistical language, Figure 1 is referred to as a "scatterplot." Typically, the degree of agreement between the raters is summarized in a statistic called a "correlation coefficient." Correlation coefficients can range from zero (indicating no agreement between raters) to one (indicating perfect agreement). For this scatterplot, the correlation coefficient (in this case, referred to as the "interrater reliability") is .75.

As evidenced by Figure 1, there clearly was a substantial amount of agreement between raters. However, this is not a high enough correlation to assume that the scores assigned by one rater would be closely matched by a second rater. While the two raters agreed closely on the vast majority of cases, there were too many students on whom raters disagreed to use the speaking instrument with just one rater rating all the students.

This level of agreement is consistent with the result observed in the 1980-81 study. In that study, however, there was a tendency for higher field correlations to be attained by raters who matched up well in ratings at the workshop. Thus, the data from the workshop examination were examined to determine whether the amount of agreement between raters could be improved by selecting those pairs that matched up well on that exam.

B. Efforts to Use the Workshop Examination Data

One of the findings of last year's study of the speaking assessment instrument was that there appeared to be a tendency for raters who later were going to correlate well in the field to match up well on ratings during the workshop. However, this finding was based on just seven pair of raters and their ratings of just four students at the end of the workshop. Consequently, a major efort in this year's study was to develop an examination to be used in the workshop.

A total of 24 student responses were recorded and played for the participants -- 12 on audiotape and 12 on videotape. Each student responded to just one speaking task, but all four tasks were presented three times each on each tape. Participants recorded their responses on a special answer sheet developed for the test.

Initial analyses of the workshop examination data showed that the ratings of the 24 students by the participants exhibited substantial agreement. Almost all participants agreed on who the best and worst speakers were. Of more concern than the overall result, however, was the identification of those pairs of raters whose ratings of the workshop students matched closely and those pairs who indicated substantial disagreement.

There was a substantial range of agreement among the pairs. While most matched up well, some pairs were in substantial disagreement, not only from each other but also from the averages of the group as a whole. Summary statistics on the workshop data are provided in Appendix C. Table C1 provides several descriptive statistics on the performance of each pair of raters; Table C2 provides similar statistics on individual raters.

Unfortunately, contrary to the expectations we had before beginning the study, there was not a strong relationship between how well pairs matched up in the workshop and how well they matched up in the field. Not to say there was no agreement -- there certainly was enough to arouse curiosity. For example, one pair (the participants from school 16) exhibited differences at the workshop that were clearly distinguishable. One rater provided extreme ratings--among the most variable of all the participants. This rater's partner was the least variable of all raters in the workshop. The second rater rated all the students within a few points of each other, even though the other raters saw substantial differences among the students being rated.

In the field, these raters performed just as they had at the workshop. While their scores of average students tended to agree, they were far apart on those who were not average. When a student was better than average, the first rater would assign extremely high scores, while the second would provide ratings just a little above average. For students below average, the first rater gave very low ratings and the second rater, a little below average. As a result, this pair had the greatest average disagreement of all pairs. On the average, the ratings they assigned were more than ten points apart.

As noted above, the expectation was that the workshop exam would predict most such cases. However, that was not the case. There were situations where raters agreed well on the workshop exam, but not in the field; and others where there was substantial disagreement on the exam, but the field ratings matched up well. While there was some relationship between performance on the workshop and performance in the field, the relationship was not high enough to accurately predict the pairs that would agree well in the field. Thus, the workshop exam did not affect the findings of the study to the degree expected.

C. Bias

A major concern of this year's study was the issue of bias. It was addressed in two parts: an examination of possible bias on the part of the raters, and an examination of possible bias in the field instrument. Both issues were investigated in terms of interaction; i.e., if two different groups scored a certain number of points apart on one occasion, did they score the same number of points apart on all other occasions?

The first examination of bias focused on the raters themselves, and their ratings of the students on the workshop exams. As mentioned earlier, the exam was in two parts: an audiotape and a videotape. Each tape included 6 students of each sex and 4 students of each ethnic group.

Appendix A includes 6 tables which provide evidence on the lack of bias in the raters and the instrument. Tables A1 through A4 include data from the workshop exam, with the last two tables providing results from the field data.

Tables A1-A4 report the average for each group of raters on each group of students. They answer the question, "Given the overall average for a particular group of students, how much higher or lower than average was that group rated by a particular set of raters?" Ideally, all entries in those four tables would be 0.0. That would mean that each group of raters arrived at the same average score for each group of students. Such an ideal cannot be expected in practice, however, because the ratings are not infallible.

None of the results in Tables A1, A3 or A4 are statistically significant. That is, the values are so close to 0.0 that they can be considered to be the result of random fluctuations due to the small sample size. However, the differences between white raters and Hispanic raters in rating black and Hispanic students, and the differences between black and white raters in rating white students, as reported in Table A2, are statistically significant. While statistically significant, these differences are small and may well be a function of the specific people who participated in the workshop.

It is interesting to note that the significant differences occurred in the audiotapes, where it had to be more difficult to detect the ethnic group of the student being rated, than on the videotape. It also is worthy of note in Tables A1 and A2 that in every case but one, the white raters assigned lower scores than did the minority raters.

The results in Tables A5 and A6 indicate that there is virtually no bias in the instrument. All groups had virtually equal scores in all scoring dimensions, although for all groups, scores were highest for delivery and lowest for content.

Scores varied a great deal from task to task, but they varied similarly for all groups. For all groups, scores were highest for the sequence task, next highest for emergency, then followed by persuasion and description. It had been the observation of teachers in the debriefing that sequence was the easiest task for students. It would seem logical to make that the first task for all speaking assessment activities.

D. Revising the Workshop and Instrument

At the conclusion of the workshop, an evaluation form was provided to all participants, and 45 participants completed and returned the form. The responses of these participants form one source of information for plans for revision.

Of the 45 respondents, 32 (71%) rated the workshop overall as "excellent", and the remaining 13 (29%) rated it as "quite good" (these were the top two categories in a scale from 1 to 5). Without question, the workshop was highly successful, provided participants with a very

positive attitude toward assessing speaking, and left them with a feeling that they were much better prepared to assess speaking than they were before attending the workshop.

In spite of their very positive response to the workshop overall, they did have some comments that would help to improve it. The first set of comments related to time. Despite the fact that the workshop was expanded from one day to two, there was not enough time to cover the material as scheduled. This problem was noted by several participants. Other suggestions for improvement made by the participants related to the repetition of the instruction, the selection of tapes that were used as exemplars, the crowded condition of the training room that made small group discussion difficult, and the failure to rotate small groups so that participants from all groups could interact.

The project staff observed the same problems and agreed with the participants, both in their general evaluation and their specific comments. One additional observation made by a member of the Department who participated in the workshop was that participants often were not ready to listen when a tape was played, and that a tone preceding each cut would have helped to prepare them for the coming example. Solutions to each of the problems noted by the participants have been proposed by the project staff; many have already been implemented in revising the project materials. These changes will be discussed in the next section of the report.

Participants also attended a debriefing held less than two months after the workshop. At the debriefing, participants were provided with tentative results of the project and through a structured interview conducted in small group settings, provided feedback on the project. Again, participants indicated great satisfaction with the workshop in general. However, they had several insights gained from their field testing that would improve the workshop.

One suggestion concerned the examples that were provided during the the workshop. Most of the examples played were of high school students, whereas the participants had to rate eighth graders. As a consequence, some raters modified their standards as they proceeded through the field testing, but others did not. Project staff felt that this might have been the major source of error and confusion in the project results. At the debriefing, participants indicated that it would be wise to focus on eighth graders in the training tapes. Somewhat to our surprise, however, they did not feel that all the tapes should be of eighth graders; the tapes should just be more focused on that age group. Consequently, the workshop tapes are being completely redone--most of the students recorded will be eighth graders.

There was a feeling that the instruction was a little more repetitive than necessary, but that there was not enough practice in scoring all four dimensions at one time. One suggestion was to teach the scoring

of delivery and language just once (since the skills of scoring those dimensions are transferable across the different speaking tasks) and then to teach the scoring of organization and content separately for each speaking task. That way, additional time would be left for practicing scoring all four dimensions at one time. The suggestion also is being incorporated into revisions of the workshop.

Participants were largely satisfied with the manual, but had some suggestions for improving that also. The principal ones dealt with the directions the administrator had to read to the student. Those directions were judged to be somewhat awkward in their wording and too detailed. An additional suggestion was to summarize the directions on one page and insert that page before the detailed directions.

Participants had several suggestions for improving the speaking tasks themselves. The participants felt the tasks could be made more applicable to eighth grade students. In particular, the emergency task was considered weak because it discouraged much speaking. The task required students to report on the telephone a fire in their house. Many students, logically enough, have been taught to make such calls as short as possible. Therefore, judging content and organization of such responses is difficult. Another set of comments focused on the persuasion task, which required students to ask the principal of their school to make a change. The participants felt that most eighth graders do not see themselves as being in a position to address the principal on the matter of a complaint, and consequently were not well equipped to handle such a task. The speaking tasks are being revised with these suggestions in mind.

Some participants proposed that the order of the tasks be changed. They felt that the sequence task should be the first one presented. Many students felt more comfortable providing a lot of detail on that item than the others. In fact, the scores on the sequence task were the highest of all the tasks. Participants also felt that the persuasion task, which would be most difficult even if revised, should remain last.

For each task, prompts were provided to get students talking if they could not think of a response on their own. Some participants felt that the prompts could have been improved by reviewing the responses of the students. In fact, many raters developed their own list of prompts after hearing the responses of the first few students. Since all sessions were recorded and the tapes returned to the project staff, this suggestion is feasible.

Participants found that it was frustrating to not be allowed to discuss each student after rating. As noted earlier, the project staff had been very insistent that no discussions of ratings take place until all ratings were concluded so that we could be assured that the degree of agreement between raters was not being improved by such interaction. To the best of our knowledge, participants accepted the frustration and conducted the ratings as requested.

Because the workshop materials established a somewhat unrealistic set of scoring standards, some participants found that they tended to assign higher scores as they went along. The adjustment from the examples at the workshop (most of high school students) to actual eighth graders required some raters to alter their scales. After the second administration of the exam at the debriefing, some participants were surprised to see how much their scores had risen from the first administration at the workshop. To counteract this problem, project staff will be developing a set of "calibration tapes" as part of the workshop revisions currently underway. The calibration tapes would be provided to participants at the end of the workshop. Then, every so often during the scoring process, they would check themselves against these tapes. If they were developing a tendency to score students more easily, or failing to recognize some characteristic of speech they had been trained to include in their ratings, they could recognize that fact and correct (or "recalibrate") themselves before scoring too many students.

Other than the above, suggestions were rather scattered and not supported by most participants. For example, some participants felt that raters should only rate for half a day. They felt that the combination of fatigue and logistics made efforts beyond a half day less productive. However, this concern was not generally voiced, and it may be a problem that can be handled on a school-by-school basis.

E. Estimates of Cost

The principal costs involved in speaking assessment are the labor costs. Material costs play such an insignificant role that they may be ignored for all practical purposes.

Raters will have to attend a workshop that probably will last two days and require some periodic brushing up of their skills. At the debriefing, raters reported no need for further training to resume rating, so it may be assumed that the periodic brushing up might be both infrequent and quick. The calibration tapes mentioned above could serve such a function.

Project participants were asked to rate 25 students per day. Most were able to do this, although most found that the administrative directions needed to be revised to accomplish this goal. As written, the directions required individual orientation of each student. Many administrators conducted small group or large group orientations. These apparently were effective and saved a substantial amount of time.

Another significant time saver was a system for having students waiting to be rated. One school established a waiting room that was monitored by Honor Society students who would otherwise be in study hall. This way, rating of the next student could begin quickly after the first was finished.

Another factor that slowed down the rating in the study was that only selected students were to be tested. Consequently, administrators had to find and withdraw from class each individual student. If the entire eighth grade were to be tested, it would be much easier to rate more students per day.

In summary, it would appear that an efficient rater should be able to test and rate at least 6 students per hour.

Preparation time is another factor in estimating time costs. Estimates of preparation time for this study ranged from 1 hour to 4 days. Obviously, preparations differed greatly from school to school, and at least one participant felt that discussion of this topic deserved much more attention in the manual. Schools differed on sending permission letters home (most did not), on who scheduled the students (usually it was up to the raters, but in some cases secretarial help was provided), and facilities and other support provided by the principal. What preparation costs would be if testing were mandated and administered to an entire grade are difficult to figure, but if the experience of this group was typical, it would be approximately 1 day of secretarial time per 100 students tested.

Another possible cost centers around the observation of several raters that it was difficult to be both an administrator and a rater. The administrator had to welcome the students, place them at ease, watch them as they were responding to questions, and operate the tape recorder. Accomplishing these tasks while also rating was felt by many to be too much for one person.

The nature of the administration and scoring tasks involved in assessing speaking are substantially different from those related to assessing reading, mathematics, and writing. The reading and mathematics tests are group-administered and the results are machine-scored. Writing also can be group-administered, but scoring must be done by hand. Speaking on the other hand, is individually administered, but the scoring is built in to the administration process.

Because of these substantial differences, relative costs are bound to be rough estimates. However, for scoring and administration, an idea of relative figures for discussion purposes are provided in Table 1.

TABLE 1
Relative Per Pupil Costs for Four Basic Skills Assessment Areas

	Reading or Mathematics	Writing	Speaking
Administration	2 minutes	2 minutes	) ten) minutes) for both
Scoring	$.35	5 minutes	

The estimates for reading and math are based on the assumption that those tests typically will take one person one hour to administer to 30 students and that computerized scoring service will provide a basic scoring package around $.35 per student. The administration time for writing

will be about the same as for reading and math, but scoring of a letter and an essay by two raters each results in scoring costs that are substantially more than those for reading and math.

These estimates would place the administration and scoring costs for assessing a student in speaking at approximately three minutes per student greater than those for writing. However, expenses for assessing speaking would be considerably less than the other areas. Two examples would be the cost of test booklets (test booklets for reading and math cost around $.50 for each area; for writing, around $.20; but nothing for speaking) and answer sheets.

V. DISCUSSION

Of the questions that this study set out to address, most have been answered. Virtually all the pieces that would be needed for large-scale implementation are complete.

The revisions to the workshop were largely successful. At this point, final revisions are being made in the workshop, and a complete set of materials will be provided to the Department by midsummer. It does not appear that the workshop could be reduced to taking less than two days, but it does not seem that a period longer than two days would be necessary, even though it was not an adequate length for the one conducted for this project. Efficiencies over that one will be built in in two ways: first, the structure of the workshop will be slightly altered so that some repetition is eliminated and more time is spent practicing the skill of rating four dimensions simultaneously; and second, a great deal of time was spent this year gathering judgments on ratings to insure that project participants, and not project staff, were establishing standards. Now that the judgments of that group have been gathered, the workshop will take on a more assertive role (i.e., rather than asking "How would you rate this student?", the workshop materials will say "This student is a 3.").

The tapes played at the workshop are being completely redone. The set used this year were gathered on students who are too old and the maturity of their speech is an unrealistic standard to set for eighth graders. This new set of tapes should make it easy for raters to maintain a consistent scoring pattern after being trained.

The idea of providing calibration tapes appears to be an excellent one. They would be helpful to raters to be able to review and remind themselves of the characteristics of different qualities of speech. These tapes have been added to the list of deliverable products for this project, and the Department will have them by midsummer.

The administrator's manual is undergoing minor revisions. The directions for students will be shortened and summarized. Additional information on how to prepare for administration will be added.

Although the speaking tasks have been previously approved, they could use revision and expansion. When speaking assessment is implemented statewide, it will not suffice to have just one set of speaking tasks. Unless alternative equivalent tasks are added, a school will not be long into testing before the four tasks are known schoolwide. Besides placing students tested earlier at a disadvantage, such a limited repertoire would lead the test to become a test of public speaking, not spontaneous speech. It will be necessary to develop alternative tasks and field test them. The workshop is being revised to include a variety of tasks as examples.

Bias appears to be less of an issue than it might have been. Certainly there was no sex bias in either the ratings or the instrument and no ethnic bias in the instrument. There may be some small amount of bias in the ratings, however. This was noticed both in the workshop results and in the comments of some raters. However, the bias that was reported by white raters was that they felt they were favoring students of ethnic minorities while the bias, if any, in the workshop results indicated a slight tendency for white raters to rate students of ethnic minorities lower. The difference between their perceptions and their scores is hard to resolve. In any case, the amount of bias observed was small relative to the differences that normally occur between judges anyway regardless of race or sex.

Critical Technical Issues

While all the issues related to development of products needed for assessing speaking will be accomplished by the conclusion of this project, the level of agreement between raters in this study was not high enough to permit all students to be rated by one rater only. It must be kept in mind that the conditions set forth in this year's study were severe, and that the reliabilities attained in a large-scale effort would be expected to be higher. Some of the conditions that helped to introduce less agreement between raters than might occur in a more general future study are:

- Raters had no feedback at all once they began rating, and some changed their standards after scoring some students. The amount of improvement in agreement that will be created by revising the workshop tapes and by providing calibration tapes might be substantial.

- Raters were recruited from a wide variety of backgrounds. If raters were selected who had more familiarity with eighth grade students, some improvement in rater agreement might be noted.

- The students selected for this study also were extreme. Students who normally would not be assessed (e.g., students who speak little English) were tested and rated. Assigning scores to these students was difficult for raters and contributed to the lack of agreement.

- The administration procedures have been refined as a result of feedback from this year's participants. Lack of standardization in administration certainly led to some lack of consistency among raters.

While it is encouraging to note that some improvement in consistency would be expected in another study, the fact remains that the level of reliability observed this year was too low to use just one rater to measure all students. Also, despite the high expectations held for it before starting the study, the workshop exam did not predict field reliability

well enough to be used to eliminate raters (although it still serves useful functions: it did identify the worst matching pair in advance, should serve as a useful device to get raters talking about their ratings, should permit especially poor raters to identify themselves and require them to recognize the disparity between their ratings and those of other workshop participants, and give confidence to successful workshop participants that they are prepared to rate students).

Consistency of Mastery Decisions. The question of how much consistency is needed between raters is certainly a question open for legitimate differences of opinion. Another way of examining consistency is to look at it in terms of how the results of the test will be used. In practice, districts will be using this test to help make pass/fail decisions about students. The results of this year's study can provide some insight into this situation.

Appendix B contains two tables that include information directly bearing on the issue. Table B1 displays the percentage of students in this year's study who would have passed any given cut score and the percentage of agreement between raters. An example may be of help in reading this table. Suppose all the data collected this year had been from one district, and suppose further that that district had selected a score of 24 as the minimum score required to pass the test. In that case, 97.2 percent of the students would have passed on the basis of one rater's scoring; and if that rating were conducted by another rater, the second rater would have arrived at the same pass/fail result for 96.5 percent of those tested.

The results reported in Table 1 demonstrate that even in the worst cases, raters agreed on almost 4 out of 5 students. But these cases occurred where the cut score was set so high that over one-third of the students failed to meet it. Based on the reported results in other Basic Skills content areas, it is more realistic to expect that most districts would not set such a difficult standard for their students. Establishment of a cut score that would be surpassed by more students would mean that a higher percentage of agreements would be achieved.

Table B2 provides additional information on the extent of agreement between pairs of raters. Over half of the ratings were within 5 points of each other; almost 3 out of 4 were within 8 points.

Table B2 provides direct information on the value of repeated assessments on students whose ratings are close to the cut score. For example, suppose a student has been rated as being 8 points above a district's cut score. It is logical to review whether it is worthwhile conducting a second assessment to see whether the pass/fail decision on that student would be altered by a second rating.

The odds that the second rater's total score for that student would be more than 8 points away from the first rater's score is around 1 out

of 4. But half of those ratings that will be more than 8 points away will be higher; thus, the odds that the second rater will rate the student below the cut score are around 1 in 8. That is, if one rater has judged a student to be 8 points above the cut score set by the district, the odds are 7 in 8 that a second rater also would pass the student.

The likelihood that scoring by a second rater will alter the judgment about a student becomes even more extreme if the district decides that the average of two ratings will be used to determine whether a student has passed or failed. Thus, suppose again that one rater has pased a student by 8 points. In order for a student to fail on the basis of the average of two ratings, the second rater would have to rate the student more than 16 points lower than the first rater. From Table B2, the likelihood of that occurring is 1/2 X 4.7%, or less than 1 case in 40.

When a student has been rated by one rater as being near the district's cut score, a valuable source of additional information may well be an assessment of that student by a second rater. However, once the student's score begins to move away from the cut score, the value of a second assessment becomes less because of the unlikelihood that the second assessment will alter the decision about that student. It may well be that a second assessment is worthwhile only if a student scores fewer than 8 points away from the cut score.

Table B3 provides the number of students in this year's study who scored fewer than 8 points away from any given cut score. The results show that upwards of 50 percent of the students tested this year were within 8 points of cut scores in the range of 42-50. For cut scores outside that range, fewer than 50 percent of the students tested would benefit from reassessment.

Use of Multiple Raters. A related issue is the recognition that equal accuracy of measurement is not needed for all students. Again, suppose our hypothetical district had set a cut score of 32 on the test. From Figure 1 (page #14), we can see that only 2 students who had a score of 48 or higher from one rater were assigned a score below 32 by the other rater. Thus, if we had had each student assessed by just one rater and then passed each student who had a score of 48 or more without further assessment, we would have made a decision agreed to by the second rater in all but 2 cases. For such students, who compose about half of the population tested, a second assessment would be redundant.

For about another quarter of the students, assessment by a second rater would result in an equally clear pass/fail decision. It is only the remaining group, perhaps one quarter of the entire population, that would require more careful assessment (perhaps by another pair of raters). Such a system would result in an average of 2 ratings per student (1 rating for half of them, 2 ratings for another quarter, and 4 ratings for the remain-

ing quarter), would have maximum precision where it is needed most (near the cut score), and would result in few misclassifications of students, even if the degree of consistency between raters found this year could not be improved upon by the improved materials and altered conditions noted above. Also, by including a variety of judges when multiple judgments were to be made, such a system would virtually eliminate any bias if it were present.

Remaining Sources of Unreliability. Clearly, the major source of inconsistency to be concerned about in assessing speaking is differences between raters--the source that was studied in depth this year. However, other smaller reasons why a student might pass one time and fail another remain univestigated. Even if raters were perfectly consistent among each other, a student might fail one time and pass another if the speaking tasks were changed or if the administrator were changed. A student might switch from one side to the other even if everything else were held constant simply because of variations in the student's own effectiveness of speech over occasions. As noted above, one effective way to reduce these uncertainties is to make multiple measurements: use more raters, more tasks, more administrators or give the student the test over more occasions.

At this point, we know or can estimate the improvement in precision that would be created by using more raters: we do not know what the effects of increasing the other factors would be. Answering the question about these sources would be a logical final step to take before recommending large-scale implementation of the speaking instrument.

Efficiencies of Individual Administration. While individually administered tests have many inefficiencies relative to group-administered tests, individually administered tests have some advantages that should not be overlooked. In contrast to the reading, writing, mathematics and listening tests, the speaking test can be administered at a time mutually convenient for both rater and student. All assessments, do not need to be conducted the same day; in fact, they might reasonably stretch over a protracted period of time. It would not be necessary, or even advisable, for raters to give up their regular classroom duties for a week or two while trying to schedule and test every student during that period. Assessments might be done over a period of several weeks, with students tested during non-class time, such as during a study hall. A rater might assess four or five students each day during their study hall; eventually, all the students in that group would be assessed. It is important to not try to make speaking assessment fit the pattern established for the other content areas. It is, by its very nature, a very different type of assessment with its own advantages and disadvantages.

A Proposed Model. Given the results acquired from this year's study, the following model would provide accurate results for any district needing individual assessments of speaking skills.

1. Have raters trained. This step would require that at least two raters from each district attend a two-day workshop to receive training. Training of more than two raters would allow the burden of assessment to be shared more equally among school staff, but would not increase the number of assessments administered to any given student. Raters ideally would be familiar with the school and its students, but would not necessarily be the teachers of the students to be assessed.

2. Assess each student once. After training, each student would be be assessed by one rater. In conducting this assessment, schools would be well advised to use the flexibility of individual assessment to its fullest advantage. Thus, assessments should be scheduled at times mutually convenient to the rater and the student. Thus, it may be a period of several weeks between the beginning of assessment activities and their conclusion.

3. Establish a cut score. After each student has been assessed, but before any review of the data, a team of people associated with the district would determine the score a student should achieve before being adjudged minimally competent in speaking.

4. Place students in one of three categories. Based on the distance between their ratings, and the cut score, students would be placed in one of three categories: pass, fail, or more information required. While the distance between the rating and the cut score may vary from district to district, 8 points seemed like a justifiable value in this year's study. As noted earlier, this value may be too high; it might be lowered for any of the reasons noted on page 24.

5. Obtain additional information on students who have not definitely passed or failed. For students near the cut score, additional information should be used to place them. One source of additional information should be a second assessment; this source alone might clarify the placement of the student. Other sources, such as teacher, counselor, principal and parent observations, could be incorporated into the decision-making process.

VI. CONCLUSIONS AND RECOMMENDATIONS

As noted in the beginning of Chapter II of this report, the development of each test designed to measure an area of the Basic Skills objectives has followed a similar pattern. Assessment of speaking skills has been on a somewhat deferred timetable because it is a content area in which little previous research had been done on which the Department could draw. Now, however, all the necessary information has been gathered to move into the final stages of implementation.

The results of this year's study have moved the development of a feasible speaking instrument a long way down the road to completion. Two major objectives were clearly achieved: all the materials necessary to implement a speaking assessment--workshop, manuals, extra tapes to keep raters on target, efficient administration procedures, etc.--will be in finished form by the conclusion of this year's effort; and concerns about bias in the raters and the instrument itself have been laid to rest. This year's study also provided an acid test of the reliability of the instrument. While higher reliability always would be desirable, the raters exhibited sufficient reliability to form definitive judgments about 50-75 percent of all tested students (depending on where a district were to set its cut score) from just one rating. The anticipated improvement to this system--the elimination of unreliable raters on the basis of a workshop exam-- was not successful, but the agreement among all raters, regardless of exam score, was sufficiently high to achieve the result noted above.

At this point, the Department is ready to complete the final two steps in the implementation of assessment of speaking skills: an investation of feasibility issues, and wide-scale implementation of the instrument and alternative forms. These final activities could be completed before the State Board meeting in June, 1983, so that the Board could take final action on the implementation of assessment of speaking skills before the 1983-84 school year begins.

Efforts to complete the final two steps would involve the following activities:

- Identification of volunteer schools and participants. eight schools, representing the range of schools in Massachusetts, would be selected from those who have volunteered to become involved in this final phase of implementation. Each school would provide four raters for training.

- Training of participants. The 32 participants would be trained to rate speaking skills using the revised, improved version of the workshop and its ancillary materials.

- Assessment of students. In each school, all eighth grade students would be assessed. A sample of 50 students would be assessed twice--once by each pair of raters. Some students would be tested with the same tasks both times; others would receive an alternate form of the test the second time. Raters would keep records of the activities they conducted and the time required to complete each task.

- Analysis of data. Two sets of analyses would be conducted. The first would be a complete study of all the sources of variability that affect reliability. The results of such a study would define for districts what the consequences of various assessment configurations would be: the benefits expected to be derived from administering an alternate form to students, from conducting a second assessment for borderline students, etc. The second analysis would be related to costs, and the most inexpensive and convenient way to conduct the assesments. It is clear from this year's study that the method chosen to conduct assessments can greatly affect the level of effort required to implement them. Such an analysis would ensure that all districts would have available to them the most facilitative processes for assessing their students.

- Large-scale production of final materials. At the conclusion of the study, sufficient materials for statewide implementation of speaking would be produced. The master manual would include information on reliability and implementation that would permit districts to select a process for assessing speaking that would involve minimum investment, intrusion, and cost.

Completion of the feasibility analysis and production of final materials is expected to require approximately 120 person-days and $10,000 worth of materials production. Therefore, producing final materials and preparing the state for complete implementation is estimated to cost approximately $40,000.

APPENDIX A

Study Results Related to Sex and Ethnic Group Differences in Ratings

Table A1

Deviations of unweighted average group ratings reported by ethnic group of students rated, for videotape workshop exam

Ethnic Group of Rater	Number of Raters	Black	Hispanic	White
Black	13	0.2	0.2	0.3
Hispanic	10	0.1	0.0	-0.3
White	28	-0.2	-0.3	0.0

Table A2

Deviations of unweighted average group ratings reported by ethnic group of students rated, for audiotape workshop exam

Ethnic Group of Rater	Number of Raters	Black	Hispanic	White
Black	13	0.0	-0.1	0.4
Hispanic	10	0.6	0.7	0.0
White	27	-0.6	-0.6	-0.4

Table A3

Deviations of unweighted average group ratings reported by sex of student rated, for videotape workshop exam

Sex of Rater	Number of Raters	Male	Female
Male	24	0.1	0.2
Female	27	-0.1	-0.2

Table A4

Deviations of unweighted average group ratings
reported by sex of student rated,
for audiotape workshop exam

Sex of Rater	Number of Raters	Male	Female
Male	23	-0.1	-0.2
Female	27	0.1	0.2

Table A5

Deviation from Group Mean On Field Ratings,
Reported By Scoring Dimension

	Blacks	Hispanics	Whites	Males	Females
Delivery	+0.1	+0.2	+0.1	+0.1	+0.1
Language	-0.1	+0.1	0.0	0.0	0.0
Organization	+0.1	-0.1	0.0	+0.1	0.0
Content	-0.1	-0.2	-0.1	-0.1	-0.1

Table A6

Deviation from Group Mean On Field Ratings,
Reported By Scoring Dimension

	Blacks	Hispanics	Whites	Males	Females
Description	-0.8	-0.6	-0.7	-0.5	-0.8
Emergency	+0.3	+0.2	+0.2	+0.3	+0.2
Sequence	+1.0	+0.7	+0.8	+0.7	+0.9
Persuasion	-0.5	-0.4	-0.3	-0.5	-0.3

APPENDIX B

Study Results Related to Agreement of Raters in the Field

TABLE B1

Percentage of Students in Study Scoring at or Above Several Given Possible Cut Scores and the Percentage of Times Both Raters Rated Students on the Same Side of the Cut Score

Cut Score	Percentage At or Above Given Cut Score	Percentage of Agreements
24	97.2	96.5
25	96.8	96.2
26	96.6	95.9
27	96.1	95.4
28	95.4	94.6
29	94.6	93.4
30	93.6	92.9
31	92.9	92.7
32	91.7	91.7
33	90.3	90.2
34	89.1	89.1
35	87.5	87.9
36	85.7	86.1
37	83.6	85.3
38	80.8	84.0
39	78.4	83.1
40	76.2	83.1
41	73.1	83.1
42	69.6	81.5
43	65.7	79.0
44	61.8	78.6
45	58.2	79.6
46	53.9	79.7
47	50.1	79.9
48	45.7	80.3
49	41.5	79.7
50	37.1	79.6
51	33.6	80.9
52	30.6	81.4
53	27.8	82.6
54	25.0	82.9

TABLE B2

Percentage of Times the Ratings Assigned to Student by Two Raters Were Within a Given Number of Points of Each Other

No. of Points Apart	Percentage of Cases	Cumulative Percentage
0	6.5	6.5
1	10.5	17.0
2	10.7	27.7
3	8.5	36.2
4	10.3	46.5
5	7.6	54.1
6	7.4	61.5
7	6.7	68.2
8	5.3	73.5
9	4.4	77.9
10	4.1	82.0
11	3.2	85.2
12	2.5	87.6
13	2.3	89.9
14	2.0	91.9
15	1.6	93.5
16	1.7	95.3
More Than 16	4.7	100.0

TABLE B3

Percentage of Students in Study Whose Ratings
Were Fewer Than Eight Points From Given Cut Scores

Given Cut Score	Percentage of Students Fewer than Eight Points From Given Cut Score
24	11.2
25	11.7
26	12.9
27	14.4
28	16.3
29	17.9
30	19.4
31	21.0
32	22.7
33	25.4
34	28.4
35	31.2
36	33.4
37	36.4
38	39.2
39	41.8
40	44.5
41	47.3
42	50.1
43	51.8
44	52.4
45	52.7
46	52.7
47	52.5
48	51.7
49	51.4
50	50.5
51	47.9
52	45.7
53	43.6
54	41.4

APPENDIX C

Statistics Describing the Performance of Raters in the Workshop and in the Field

TABLE C1

Statistics Describing the Performance of Pairs of Raters

	WORKSHOP RESULTS						FIELD RESULTS	
	Audio Test			Video Test				
School	Mean Difference	Total Discrepancy	Correlation	Mean Difference	Total Discrepancy	Correlation	Mean Discrepancy	Correlation
1	-.2	14	.93	-.2	21	.78	3.8	.84
2	-	-	-	-.5	30	.61	7.0	.72
3	-.1	23	.85	1.2	13	.70	5.3	.80
4A	.1	34	.58	1.7	43	.11	8.8	.77
4B	1.0	40	.62	.8	33	.58	6.7	.87
4C	1.3	24	.78	1.1	33	.40	6.6	.77
5A	.1	34	.58	1.7	43	.11	8.8	.77
5B	-.5	12	.96	-.7	20	.68	4.6	.75
5C	-.6	36	.56	-2.4	35	.48	7.7	.84
6	-.1	22	.78	.0	20	.27	7.6	.86
7	1.0	31	.71	-.4	27	.72	5.7	.86
8	-.4	20	.82	-.2	10	.81	2.3	.77
9	-2.2	27	.91	.6	36	.40	6.2	.89
10A	-.5	10	.96	.6	25	.64	6.2	.79
10B	1.1	25	.79	-.6	26	.69	5.1	.85
10C	1.6	25	.84	-1.2	23	.71	5.2	.79
11	-2.4	35	.79	-.3	18	.88	6.7	.73
12A	1.2	23	.85	1.6	25	.77	4.1	.81
12B	1.1	21	.93	-1.1	26	.78	4.7	.79
12C	-.1	22	.92	-2.7	33	.86	5.1	.75
13A	1.1	29	.78	.0	30	.60	8.7	.85
13B	1.4	18	.92	-.1	17	.87	7.2	.75
13C	.3	17	.90	-.1	23	.79	5.9	.83
14	.2	22	.86	-1.4	31	.63	-	-
15	2.3	36	.66	-3.6	43	.74	7.0	.72
16	.5	35	.70	.4	39	.50	10.4	.72
17	.4	17	.90	-.2	20	.78	5.7	.78
18	-1.4	33	.69	-.3	18	.77	6.3	.82
19	.9	19	.87	.0	14	.89	3.6	.86
20	.5	12	.89	1.3	20	.86	8.3	.54
21	2.1	27	.94	1.8	21	.79	4.6	.66
22	.4	19	.90	.7	32	.74	6.4	.62

TABLE C2

Statistics Describing the Workshop Performance of Individual Raters

	Audio Test			Video Test		
Rater	Reliability	Mean	Standard Deviation	Reliability	Mean	Standard Deviation
1A	.92	9.9	4.2	.92	10.5	3.5
1B	.93	10.1	3.9	.85	10.7	3.0
2A	.93	10.8	3.9	.77	11.4	3.8
2B	-	-	-	.85	11.9	3.1
3A	.90	11.7	4.3	.84	13.0	3.1
3B	.93	11.8	3.2	.89	11.8	3.4
4A	.88	11.5	5.1	.81	11.8	4.2
4B	.85	11.8	2.5	.82	12.1	2.3
4C	.74	10.5	3.5	.74	11.0	3.1
5A	.93	10.3	3.6	.85	11.4	2.8
5B	.78	10.2	4.0	.33	9.7	3.1
5C	.93	10.8	3.8	.90	12.1	2.6
6A	.79	12.3	2.8	.96	12.3	3.7
6B	.98	12.4	3.8	.81	12.3	2.8
7A	.83	11.8	4.3	.53	10.2	3.9
7B	.91	10.8	3.2	.87	10.6	2.0
8A	.83	10.9	2.3	.73	11.0	1.6
8B	.96	11.3	3.5	.84	11.2	1.0
9A	.92	9.1	4.5	.86	11.3	4.2
9B	.99	11.3	3.8	.55	10.7	2.2
10A	.94	11.4	4.4	.83	11.2	3.5
10B	.97	11.9	4.3	.74	10.6	3.0
10C	.89	10.3	4.0	.94	11.8	3.0
11A	.90	9.6	3.8	.91	11.2	3.8
11B	.88	12.0	4.3	.89	11.5	3.0
12A	.97	11.9	3.3	.92	12.7	3.6
12B	.93	10.7	2.8	.86	11.1	2.7
12C	.95	10.8	4.7	.81	13.8	3.9
13A	.89	12.2	3.6	.77	11.7	3.4
13B	.94	11.1	3.9	.82	11.7	3.8
13C	.97	10.8	3.4	.83	11.8	3.0
14A	.91	10.4	4.9	.68	10.2	2.6
14B	.94	10.3	3.6	.89	11.6	3.7

TABLE C2 (Continued)

Statistics Describing the Workshop Performance of Individual Raters

	Audio Test			Video Test		
Rater	Reliability	Mean	Standard Deviation	Reliability	Mean	Standard Deviation
15A	.97	10.3	3.5	.84	8.9	2.6
15B	.72	12.6	3.6	.85	12.5	3.4
16A	.81	12.3	4.8	.84	12.3	4.6
16B	.95	11.8	2.2	.72	11.9	1.3
17A	.97	11.1	4.2	.88	11.5	3.3
17B	.93	10.7	3.8	.87	11.7	2.7
18A	.91	9.6	4.1	.91	10.8	3.3
18B	.87	11.0	3.7	.79	11.1	3.4
19A	.93	11.8	3.4	.84	12.3	3.5
19B	.94	10.9	3.6	.88	12.3	3.2
20A	.90	11.7	3.1	.83	11.6	3.3
20B	.85	11.2	2.8	.90	10.3	3.0
21A	.97	12.0	3.2	.78	12.1	2.3
21B	.97	9.9	4.1	.81	10.3	2.4
22A	.84	10.7	2.7	.85	11.2	2.8
22B	.93	10.3	4.2	.75	10.5	4.3

BASIC SKILLS IMPROVEMENT POLICY

Department Resources - Speaking

Assessment Instrument and Information

Assessment of Basic Speaking and Listening Skills; State of the Art and Recommendations for Instrument Development - Volume I Summary Review	September, 1979
Assessment of Basic Speaking and Listening Skills; State of the Art and Recommendations for Instrument Development - Volume II Test Reviews	September, 1979
Listening and Speaking Summary Report	September, 1980
Listening and Speaking Technical Report	November, 1980
Listening and Speaking Development Report	February, 1981
Listening and Speaking Overview	March, 1981
Final Report Speaking Assessment 1981	August, 1981
Development of State Speaking Assessment Instrument: Reliability and Bias Study - Summary Report	June, 1982
Development of State Speaking Assessment Instrument: Reliability and Bias Study - Technical Report	July, 1982

This report was prepared with the assistance of:

RMC Research Corporation
400 Lafayette Road
Hampton, NH 03842